Chapitre 4 : Analyse de la variance

1 Variabilité expliquée par le modèle

Dans le modèle linéaire,

$$\begin{aligned}\widehat{\mathbf{Y}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \widehat{\beta}_0\mathbf{X}_{.,0} + \widehat{\beta}_1\mathbf{X}_{.,1} + \cdots + \widehat{\beta}_p\mathbf{X}_{.,p} \in \mathrm{Im}\,\mathbf{X}.\end{aligned}$$

Sont équivalents :

(i) choisir $\widehat{\boldsymbol{\beta}}$ avec

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\beta} \left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|^2$$

(ii) choisir $\widehat{\boldsymbol{\beta}}$ tel que

$$\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \perp \mathrm{Im}\,\mathbf{X}$$

(iii) choisir $\widehat{\boldsymbol{\beta}}$ tel que $\mathbf{X}\widehat{\boldsymbol{\beta}}$ soit la projection orthogonale de $\mathbf{Y}$ sur $\mathrm{Im}\,\mathbf{X}$

Donc,

$$\mathbf{Y} = \underbrace{\mathbf{Y} - \widehat{\mathbf{Y}}}_{\in \mathrm{Im}\mathbf{X}^{\perp}} + \underbrace{\widehat{\mathbf{Y}}}_{\in \mathrm{Im}\mathbf{X}} \quad \text{et}$$
$$\|\mathbf{Y}\|^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 + \|\widehat{\mathbf{Y}}\|^2$$

avec le théorème de Pythagore.

Pour obtenir la décomposition classique de la variance, il faut travailler dans le sous-espace orthogonal à $\mathbf{1} \in \mathbb{R}^n$. Notons que $\mathbf{1} \in \mathbf{Im}\mathbf{X}$. Si on utilise la première égalité, on obtient encore

$$\mathbf{Y} - \bar{\mathbf{Y}} = \underbrace{\mathbf{Y} - \widehat{\mathbf{Y}}}_{\in \mathrm{Im}\mathbf{X}^{\perp}} + \underbrace{\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}}_{\in \mathrm{Im}\mathbf{X}}$$

où $\bar{\mathbf{Y}} = (\bar{Y}, \ldots, \bar{Y})$. Donc

$$SST = SSR + SSE, \quad \text{où}$$

- $SST = \|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2$ est la variabilité totale,
- $SSR = \|\widehat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2$ est la variabilité expliquée par le modèle et
- $SSE = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$ est la variabilité résiduelle

2 Analyse de la variance à un facteur

2.1 Le point de vue décomposition de la variance

On s'intéresse à un modèle où il n'y a qu'une seule covariable catégorielle G à $p+1$ modalités, que l'on doit remplacer par des variables binaires $X_1, \ldots, X_p$.

Il est usuel de représenter la décomposition de la variance de la façon suivante

	DF	**SS**	**MS**	**F-value**	**p-value**
G	p	SSR	SSR/p	$\frac{SSR/p}{SSE/(n-p-1)}$	*
residuals	$n-p-1$	SSE	$SSE/(n-p-1)$		
total	$n-1$	SST			

Les deux dernières colonnes n'ont de sens que sous les hypothèses gaussiennes, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Le F-test et sa p-value permet de tester si les effets en facteurs des binaires sont tous nuls ou si l'un au moins d'entre eux est non nul. On peut le voir comme un test de comparaison de $p+1$ groupes.

3 Tests de comparaison de groupes

3.1 Tests de Student

Supposons que la population est divisée en deux groupes, et que l'on s'intéresse à comparer les moyennes μ_0, μ_1 de Y au sein des deux groupes. On veut tester

$$H_0 \ : \ \mu_0 = \mu_1, \quad \text{contre} \quad H_1 \ : \ \mu_0 \neq \mu_1.$$

On dispose d'un échantillon de taille n que l'on peut renuméroter de telle sorte que $Y_{0,1}, \ldots, Y_{0,n_0}$ modélisent les observations au sein du premier groupe, et $Y_{1,1}, \ldots, Y_{1,n_1}$ au sein du second groupe, avec $n_0 + n_1 = n$.

Pour mimer les hypothèses gaussiennes et d'homoscédasticité, on suppose que

les $Y_{i,j}$ sont indépendants et que

$$Y_{0,1},\dots,Y_{0,n_0} \sim \mathcal{N}(\mu_0;\sigma^2),$$
$$Y_{1,1},\dots,Y_{1,n_1} \sim \mathcal{N}(\mu_1;\sigma^2).$$

Dans ce cas, on estime les deux moyennes et σ^2 par

$$\widehat{\mu}_0 = n_0^{-1}(Y_{0,1}+\cdots+Y_{0,n_0}),$$
$$\widehat{\mu}_1 = n_1^{-1}(Y_{1,1}+\cdots+Y_{1,n_1}),$$
$$\widehat{\sigma}^2 = \frac{1}{n-2}\left(\sum_{j=1}^{n_0}(Y_{0,j}-\widehat{\mu}_0)^2+\sum_{j=1}^{n_1}(Y_{1,j}-\widehat{\mu}_1)^2\right).$$

Il est facile de voir que $\widehat{\mu}_0$ et $\widehat{\mu}_1$ sont indépendants, et que

$$\widehat{\mu}_i \sim \mathcal{N}\left(\mu_i,\frac{\sigma^2}{n_i}\right), \quad i=0,1.$$

De plus, avec le théorème de Cochran, $\widehat{\sigma}^2$ est indépendant de $(\widehat{\mu}_0,\widehat{\mu}_1)$ et

$$(n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Pour réaliser le test, on introduit la statistique

$$T = \frac{\widehat{\mu}_0 - \widehat{\mu}_1}{\widehat{\sigma}\sqrt{n_0^{-1} + n_1^{-1}}}.$$

Sous l'hypothèse nulle H_0, les résultats précédents montrent que T suit $t(n-2)$, la loi de Student à $n-2$ degrés de liberté. Sous l'hypothèse alternative, à mesure que μ_0 s'éloigne de μ_1, $|T|$ devient grand.

On décide donc en faveur de H_0 si la valeur observée de T est dans un intervalle de la forme $[-c; c]$ et en faveur de H_1 sinon. La valeur de c est choisie pour que le risque maximal de première espèce soit égal à α, ce qui conduit à

$$c = \Phi_{n-2}^{-1}(1-\alpha/2) = \text{le quantile d'ordre } 1-(\alpha/2) \text{ de } t(n-2).$$

Et la p-value du test est

$$p(t_{\text{obs}}) = \mathbb{P}_{T\sim t(n-2)}\Big(|T| > |t_{\text{obs}}|\Big) = 2\Big(1-\Phi_{n-2}(|t_{\text{obs}}|)\Big),$$

où Φ_{n-2} est la fonction de répartition de $t(n-2)$.

Remarque. Les équations

$$Y_{0,1}, \dots, Y_{0,n_0} \sim \mathcal{N}(\mu_0; \sigma^2),$$
$$Y_{1,1}, \dots, Y_{1,n_1} \sim \mathcal{N}(\mu_1; \sigma^2).$$

reviennent à dire que

$$Y_{i,j} = \mu_i + \varepsilon_{i,j}, \quad \text{où } \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

avec indépendance des $\varepsilon_{i,j}$. Quitte à poser $\beta_0 = \mu_0$ et $\beta_1 = \mu_1 - \mu_0$, on a

$$Y_{i,j} = \beta_0 + \beta_1 \mathbf{1}\{i = 1\} + \varepsilon_{i,j}.$$

Il s'agit bien du modèle linéaire avec une seule variable explicative binaire, qui est l'indicatrice d'être dans le groupe n°1. Le t-test présenté plus haut est exactement le test de nullité de β. Dans ce cas où il n'y a qu'une seule covariable, ce t-test est équivalent au F-test (ou test de Fisher d'influence des covariables).

3.2 Le point de vue test de comparaison de plusieurs groupes

Lorsque la population est divisée en K groupes, cela veut dire qu'il existe une variable catégorielle avec K modalités (les noms des groupes). Pour l'introduire

dans le modèle linéaire, il faut donc :

— choisir une modalité de référence (numérotée 0),

— introduire autant de variables binaires qu'il y a d'autres modalités : $X_1, \ldots, X_{K-1}$.

Le modèle linéaire est donc, pour un individu $(X_1, \ldots, X_{K-1}, Y)$ pris au hasard dans la population :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \text{où } \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Si on note G le numéro du groupe auquel appartient cet individu, on a

$$\mathbb{E}\Big(Y \Big| G = i\Big) = \begin{cases} \beta_0 & \text{si } i = 0, \\ \beta_0 + \beta_i & \text{si } i = 1, \ldots, K-1. \end{cases}$$

Autrement dit, si on pose $\mu_i = \mathbb{E}\Big(Y \Big| G = i\Big)$ la moyenne de Y au sein du groupe numéro i,

Proposition 1. *On a $\beta_0 = \mu_0$ et pour $i = 1, \ldots, K-1$, on a $\beta_i = \mu_i - \mu_0$.*

Dans cette situation, si on a un jeu de données de n observations, il est courant de les re-numéroter de la façon suivante :

— $Y_{0,1}, \ldots, Y_{0,n_0}$ sont les n_0 observations du groupe 0 (groupe de référence),
— $Y_{i,1}, \ldots, Y_{i,n_i}$ sont les n_i observations du groupe i.

Le nombre total d'observations est $n = n_0 + \cdots + n_{K-1}$.

Le modèle statistique sur les observations est donc

$$Y_{0,j} \sim \mathcal{N}(\beta_0, \sigma^2), \quad Y_{i,j} \sim \mathcal{N}(\beta_0 + \beta_i, \sigma^2), \text{ si } i = 1, \ldots, K-1.$$

Proposition 2. *Soit*

$$\widehat{\mu}_i = n_i^{-1}(Y_{i,1} + \cdots + Y_{i,n_i})$$

L'estimateur du maximum de vraisemblance de $(\beta_0, \ldots, \beta_{K-1}, \sigma^2)$ *est donné par*

$$\widehat{\beta}_0 = \widehat{\mu}_0, \quad \widehat{\beta}_i = \widehat{\mu}_i - \widehat{\mu}_0 \ (i = 1, \ldots, K-1)$$

et

$$\widehat{\sigma}^2 = \frac{1}{n-K} \sum_{i=0}^{K-1} \sum_{j=1}^{n_i} (Y_{i,j} - \widehat{\mu}_i)^2.$$

L'objectif de l'analyse de la variance est de tester

$$H_0 \ : \ \mu_0 = \mu_1 = \cdots = \mu_{K-1} \quad \text{contre} \quad H_1 \ : \ \exists i \neq i', \ \ \mu_i \neq \mu_{i'}.$$

Vue la proposition 1, il s'agit du F-test d'intérêt du modèle linéaire où H_0 est $\beta_1 = \cdots = \beta_{K-1}$.

Notons que si l'on décide en faveur de H_1 avec de F-test, on ne sait pas quels sont les deux groupes qui se distinguent par leurs moyennes. Pour répondre à cette question, il faut faire autant de t-tests que de paires de groupes, ce qui soulève un problème de tests multiples qui dépasse le contenu de ce cours.

4 Au delà de l'analyse de la variance à un facteur

4.1 Le problème de la décomposition de SST

Pour examiner un modèle linéaire, on le compare à des versions simplifiées de lui-même, de manière à comprendre l'intérêt des variables dans la prédiction de Y. Nous allons donc comparer deux modèles :

— le modèle M_q, ajusté sur la matrice de design $\mathbf{X}^{(q)}$ de rang q,
— le modèle M_r, ajusté sur la matrice de design $\mathbf{X}^{(r)}$ de rang $r > q$.

On suppose que les modèles sont emboîtés, c'est-à-dire que $\mathrm{Im}\mathbf{X}^{(q)} \subset \mathrm{Im}\mathbf{X}^{(r)} \subset \mathbb{R}^n$, où n est le nombre d'observations. Cette inclusion est vraie quand, par exemple, on enlève une colonne ou un ensemble de colonne à la matrice de design $\mathbf{X}$ originale.

Pour simplifier, on va supposer que $\mathbf{1} \in \mathrm{Im}\mathbf{X}^{(q)}$.

Si on ajuste le modèle M_r, on obtient une estimation $\widehat{\boldsymbol{\beta}}^{(r)}$ et une prédiction

$$\widehat{\mathbf{Y}}^{(r)} = \mathbf{X}^{(r)}\widehat{\boldsymbol{\beta}}^{(r)} \in \mathrm{Im}\mathbf{X}^{(r)}$$

De même pour le modèle M_q, on obtient une estimation $\widehat{\boldsymbol{\beta}}^{(q)}$ et une prédiction

$$\widehat{\mathbf{Y}}^{(q)} = \mathbf{X}^{(q)}\widehat{\boldsymbol{\beta}}^{(q)} \in \mathrm{Im}\mathbf{X}^{(q)}.$$

Malheureusement, dans la décomposition ci-dessous

$$\mathbf{Y} = \widehat{\mathbf{Y}}^{(q)} + \left(\widehat{\mathbf{Y}}^{(r)} - \widehat{\mathbf{Y}}^{(q)}\right) + \left(\mathbf{Y} - \widehat{\mathbf{Y}}^{(r)}\right)$$

n'est pas la somme de trois vecteurs orthogonaux en règle générale. Pour obtenir une décomposition de SST, il faut donc travailler différemment. La théorie complète est compliquée et dépasse largement le cadre de ce que l'on peut exposer dans ce cours.

Un cas simple où ces trois vecteurs sont orthogonaux est celui où l'on ajoute à $\mathbf{X}^{(q)}$ des colonnes qui sont orthogonale à $\mathbf{Im}\mathbf{X}^{(q)}$:

$$\mathbf{X}^{(r)} = \left(\mathbf{X}^{(q)} \middle| \mathbf{X}^{(r-q)}\right).$$

Cela revient à supposer que **les variables que l'on ajoute sont décorellées des variables de** M_q.

4.2 Analyse de la variance à deux variables indépendantes

Dans cette section, on suppose s'intéresse à **deux variables catégorielles indépendantes**. Autrement dit, il y a deux partitions de population :

- une première partition en K groupes,
- une seconde partition en L groupes.

On note G la variable qui décrit le numéro du groupe (de 0 à $K-1$) de la première partition. Et H celle pour la seconde partition, de 0 à $L-1$.

L'hypothèse d'indépendance revient à dire que, pour tout $i \geq 0$, tout $j \geq 0$,

$$\mathbb{P}(G = i, H = j) = \mathbb{P}(G = i)\mathbb{P}(H = j).$$

Introduire deux partitions revient en fait à introduire une seule partition à $K \times L$ groupes, chaque groupe de cette dernière partition est alors caractérisée par la valeur de la paire (G, H).

Il est d'usage de re-numéroter les observations, et les variables aléatoires qui les modélisent, de telle sorte que

$$Y_{i,j,k}$$

est la k-ème observation appartenant aux groupes $G = i$ et $H = j$ simultanément. Le nombre de telles observations est noté $n_{i,j}$.

Si on note $\mu_{i,j} = \mathbb{E}\big(Y \big| G = i, H = j\big)$, le modèle où

$$Y_{i,j,k} \sim \mathcal{N}(\mu_{i,j}, \sigma^2), \quad k = 1, \ldots, n_{i,j}$$

correspond à un modèle linéaire utilisant les variables G, H **avec leurs interactions**. On introduit les variables binaires :

$$U_i = \mathbf{1}\{G = i\} \ (i = 1, \ldots, K-1) \quad \text{et} \quad V_j = \mathbf{1}\{H = j\} \ (j = 1, \ldots, L-1).$$

Alors,

$$Y = \mu + \sum_{i=1}^{K-1} \alpha_i U_i + \sum_{j=1}^{L-1} \beta_j V_j + \sum_{i=1}^{K-1} \sum_{j=1}^{L-1} \gamma_{i,j} U_i V_j + \varepsilon, \quad \text{où } \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Dans ce cas, on a

Proposition 3. *Pour* $i = 1, \ldots, K-1$ *et* $j = 1, \ldots, L-1$,

$$\mu_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$$

Pour $i = 1, \ldots, K-1$,

$$\mu_{i,0} = \mu + \alpha_i.$$

Pour $j = 1, \ldots, L-1$,

$$\mu_{0,j} = \mu + \beta_j.$$

Enfin

$$\mu_{0,0} = \mu.$$

Réciproquement, on a, si $i, j \geq 1$

$$\begin{cases} \mu = \mu_{0,0}, \quad \alpha_i = \mu_{i,0} - \mu_{0,0}, \quad \beta_j = \mu_{0,j} - \mu_{0,0} \\ \gamma_{i,j} = \mu_{i,j} - \mu_{i,0} - \mu_{0,j} + \mu_{0,0} \end{cases}$$

4.3 Le cas équilibré

On suppose que le nombre d'observations $n_{i,j}$ ne dépend pas de (i,j) et donc,

$$\forall i, \forall j, \quad n_{i,j} = n_i n_j$$

où $n_i > 1$ et $n_j > 1$ sont deux nombres entiers non nuls. C'est le cas le plus simple. Mais, le nombre total d'observations peut devenir rapidement grand.

On pose

$$\widehat{\mu}_{i,j} = n_{i,j}^{-1} \sum_{k=1}^{n_{i,j}} Y_{i,j,k}.$$

Proposition 4. *L'estimateur du maximum de vraisemblance du modèle avec interaction est donné par*

$$\begin{cases} \widehat{\mu} = \widehat{\mu}_{0,0}, \quad \widehat{\alpha}_i = \widehat{\mu}_{i,0} - \widehat{\mu}_{0,0}, \quad \widehat{\beta}_j = \widehat{\mu}_{0,j} - \widehat{\mu}_{0,0} \\ \widehat{\gamma}_{i,j} = \widehat{\mu}_{i,j} - \widehat{\mu}_{i,0} - \widehat{\mu}_{0,j} + \widehat{\mu}_{0,0} \end{cases}$$

et

$$\widehat{\sigma}^2 = \frac{1}{n - KL} \sum_{i=0}^{K-1} \sum_{j=0}^{L-1} \sum_{k=1}^{n_{i,j}} (Y_{i,j,k} - \widehat{\mu}_{i,j})^2.$$

Les tableaux d'analyse de la variance se présentent alors de cette façon :

	df	**Sum Sq**	**Mean Sq**	F**-value**	p**-value**
G	$K-1$	SSG	$SSG/(K-1)$		
H	$L-1$	SSH	$SSH/(L-1)$		
$G:H$	$(K-1)\times(L-1)$	$SSGH$	$SSGH/((K-1)(L-1))$		
residuals	$n-KL$	SSE	$\widehat{\sigma}^2$	—	—
total	$n-1$	SST			

Le test, sur la ligne des interactions $G:H$, permet de savoir si celles-ci sont utiles. Les tests, sur les lignes G et H, permettent de savoir si ces variables ont un effet moyen, ou si leur effet n'apparait qu'au travers des interactions. Suivant les cas pratiques, ces test ont ou n'ont pas de sens...

4.4 Plans d'expérience déséquilibrés

Ce sont typiquement des cas où l'on n'a pas pu échantillonner tous les cas $G = i$ et $H = j$, pour $i = 0, \ldots, K-1$ et $j = 0, \ldots, L-1$. Bien souvent, on ne peut pas inclure les interactions dans les modèles sur de tels jeux de données par manque d'observations.

La section 2.5 du chapitre 2 de cours « *Le Modèle Linéaire et ses Extensions* » de L. Bel, J.J. Daudin *et al.* présente le cas d'un plan en blocs incomplets au travers d'un exemple. . .

5 Analyse de la covariance

Il existe tout un bestiaire d'analyse de la variance. Voir, par exemple « *Analysis of Variance and Covariance* » de Doncaster et Davey.

Le plan du livre est disponible ici : https://www.southampton.ac.uk/ cpd/anovas/

et les exemples là :
https://www.southampton.ac.uk/ cpd/anovas/datasets/index.htm

La section 2.6 du chapitre 2 de cours « *Le Modèle Linéaire et ses Extensions* » de L. Bel, J.J. Daudin *et al.* présente un cas où les covariables sont corrélées, et non contrôlables par plan d'expérience. . . .